DOCUMENT RESUME

ED 310 151 TM 013 744

AUTHOR Forster, Fred; Karr, Chad
TITLE Things Mother Never Told You about Fit Statistics.
PUB DATE Mar 89
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Goodness of Fit; Grade 8; *Latent Trait Theory; *Sample Size
IDENTIFIERS *F Test; *Outliers

ABSTRACT

It is postulated that mean square fit statistics do not adequately evaluate the quality of test items adequately. The fit statistics arise from the desire to have statistical (computerized) procedures to assess item quality transcending the group on which they are computed. It appears that none of the current formulations meet these criteria. The limitations of current fit statistics are demonstrated. Problems identified are related to sample size and to "outliers." It is proposed that psychometricians use the plot of the item characteristic curve in place of fit statistics. Random samples of 300 high achievers, 300 low achievers, and 300 sample students were drawn from the test results of 5,000 eighth graders, who took a 65-item statewide mathematics test, in order to demonstrate the "f" coefficient--a new correlation type coefficient. It is concluded that the "f" coefficient may prove useful in analyzing item characteristic curves and assessing item quality, but is not consistent across samples or for problem items. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED310151

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CHAD KARR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THINGS MOTHER NEVER TOLD YOU ABOUT FIT STATISTICS

Fred Forster

Microsystems for Education and Business, Inc.
395 SW 6th Street
Gresham, OR 97080
(503) 666-7883

Chad Karr

Department of Psychology
Portland State University
P.O. Box 751
Portland, OR 97207

Presented at the
1989 AERA Annual Meeting
San Francisco, CA
March 28, 1989

TM013744

BEST COPY AVAILABLE

THINGS MOTHER NEVER TOLD YOU ABOUT FIT STATISTICS

The Premise

We believe mean square fit statistics do not evaluate the quality of test items adequately. The fit statistics arise from the desire to have a statistical (computerized) procedure to assess item quality transcending the group on which they are computed. We believe that none of the current formulations meet these criteria.

To explore this issue we must start with a working definition of "reality," against which to contrast the performance of the various statistics. Those of us who use the Rasch model are lucky because we can plot the actual performance of students in the same metric as the "theoretical" curve predicted by the model. Therefore, we can assess the effectiveness of a fit statistic based on its ability to describe the match between the theoretical curve and the way students actually performed.

This paper demonstrates the limitations of current fit statistics and proposes that psychometricians use the plot of the item characteristic curve in their place. It also describes a new correlation type coefficient (the $\underline{f}$ coefficient) which may prove useful in analyzing item characteristic curves and assessing item quality.

Our Experience with Fit

To put this discussion in perspective, we believe that you should try to use every available measure of item quality. In the battle to maintain accurate measurement, we need all the help we can get. In our own work we use the point-biserial, mean square fit, and a new coefficient which we call the $\underline{f}$ coefficient. But, we advise against the blind application of any of these statistics without the examination of the item characteristic curve.

Our work with the mean square out-fit statistic has uncovered two major problems. One problem relates to sample size. As the number of students in the sample increases, typically, the value of the fit statistic grows much like the $\underline{F}$ or $\underline{t}$ test with large samples.

A second problem relates to outliers. If a few of the highest scoring students on a test miss the item, or if a few of the lowest scoring students answer it correctly, the size of the out-fit balloons.

Similarly, the mean square in-fit statistic has disappointed us. While it avoids the sample size problem, it focuses on the responses for those students for whom the item was poorly matched; i.e., those students for whom the item was much too easy or much too difficult. The effect of this differential weighting of responses in calculating in-fit is shown in Table 1. When the item is well matched to the student, the contribution to the mean square in-fit statistic is approximately 1, but when the item is poorly matched (either too hard or too easy) the contribution to in-fit is either very small or very large. As a result, nearly all the variance in this statistic is due to students at the extremes of the measurement capacity of the item.

Please do not generalize these comments to the robustness of Rasch calibration itself. In numerous studies we have found that, in normal testing situations, Rasch calibrations are sample independent from group to group across grades, geographic regions, and achievement level. Our present criticism applies only to the fit statistics derived from Rasch calibration and Rasch scaling.

Our Experience with Item Banking

The work on which we base our conclusions comes from the analysis of hundreds of tests and thousands of items used to build the NWEA item banks in reading, mathematics, and language usage.

Our quality control procedures screened each item for curriculum validity and appropriateness for use in the field tests. Our population included public school students in the states of Oregon, Washington, California, and Michigan. The teachers who volunteered for this field testing received quick turn-around reports describing the performance of their students. We screened the answer sheets for missing marks and dropped the test results of students scoring above 90 percent or below 25 percent correct on the test as a whole. We checked results classroom by classroom and dropped a class if its calibrations were inconsistent, and dropped an item if its calibrations were inconsistent across groups. Using these procedures, we removed the "noisy" data (about 15 percent) before we performed the Rasch analysis. (A detailed description of these procedures is given by Forster in the Rasch Newsletter, Vol.2, No.2, Summer, 1988.) In short, we took the task of securing clean test information seriously and put in a commensurate amount of effort.

Based on this experience, we believe that other researchers doing large-scale item banking or managing other serious applications of the Rasch model to the basic skills of reading, mathematics and language will find similar results.

The Item Curve

We found that the item (characteristic) curve was the best available method to analyze test questions. Using a little ingenuity and the Rasch equations, one can plot this curve showing the agreement between the expected and the actual percent correct for groups of students at different achievement levels.

George Ingebo has suggested a valuable improvement to the item curve for samples under 1000. We "smooth" the data by combining responses across a 5 RIT interval for each observed scores. As shown in Figures 1 and 2, the smoothed item curve gives a clearer picture of the performance of an item across the range of student achievement represented in the sample.

An important advantage of the item curve is that it becomes more accurate as sample size increases. A larger sample produces more stable data points to define the characteristics of the item more precisely. A second advantage is that the item curve summarizes all the information available on the performance of a test question. It shows the level (of difficulty) of the item and the match between the item and the group taking the test. It documents what we know about the item and what we don't know; i.e., whether the item was too easy or too hard to be adequately analyzed by either this graphical method or any of the fit statistics. Most important, the item curve approaches the question of model fit directly by asking the question, "Does the actual performance of students match our expectations?" The fit statistics, which are based on the squared residuals, approach the question backwards by asking the double negative question; i.e., does the item not misfit the model? Like
F and t tests, the larger the sample, the more certain we are to conclude that there is a difference, even though that difference may be trivial.

Some test developers may feel uneasy interpreting the item curve to determine item quality rather than using a simple number like the mean square, but the validity of a statistic procedure lies in its accuracy, not its convenience. For example, the use of confidence intervals has a clear advantage over a simple F or t test. If we use samples of several thousand to find that Canadians are, on the average, 1/100 of an inch taller than people living in the U.S., and our method of measurement is accurate to 1/10 of an inch, how significant is that difference? Does this practical meaning change if the probability of an F test is .01, or .001, or .001? On the other hand the confidence interval would show the bounds on the actual size of that difference, so it could be interpreted correctly. We believe the item curve provides the same kind of corrective interpretation of these data.

An Alternative Statistic

We developed a new statistic called the f coefficient to quantify item fit directly. This statistic is a variation of the eta coefficient designed to determine the match between expected and actual components of the item curve as the square root of the ratio of the variance accounted for by the curve to the total variance. Following are the steps in calculating the f coefficient.

1. Plot the observed p for groups at different achievement levels against the logistic curve.

2. Calculate the sum of the squared residuals and divide by n:

   $$\text{Res Variance} = \sum (p_{obs} - p_{th})^2/n$$

3. Calculate the variance of the theoretical p values:

   $$\text{Variance} = \sum (p_{th} - \bar{p})^2/n$$

4. Calculate the ratio of variance accounted for to the total variance and take the square root to obtain the f coefficient:

   $$f = \sqrt{(\text{Residual Variance}/\text{Variance})}$$

Method

For this study, we used the test results from 5000 8th grade students who took a 65-item statewide mathematics test. We drew a random sample of 300 students (HIGH) from just the top half of the distribution, a second sample of 300 (LOW) from just the bottom half of the distribution, and a third sample of 300 (SAMPLE) from the entire distribution. We plotted the item curves and the point biserial, in-fit, out-fit, and f coefficient for each item using the total population and each of the samples. We then selected representative items to assess the effectiveness of each statistic in identifying the quality of the items and problems which were apparent from the analysis of the item curve.

Results

Table 2 summarizes the item statistics for 10 items which we judged to be "ideal" and 6 items which we judged to be "problem" items. (See Figure 3 for an example of a typical "ideal" item curve and Figure 4 for an example of typical "problem" item curve.)

The point biserial varied widely among the samples and was most sensitive to the restrictions in the sample variance for the HIGH and LOW groups. The in-fit mean square did not appear to be of much use in identifying the problem items. The out-fit mean square appeared to be more sensitive to item quality but did not show consistency from sample to sample. Finally, the $\underline{f}$ coefficient appeared to be of some value in assessing the quality of the ideal items, but was not consistent across samples or for the problem items.

Conclusions

Based on these data, we do not believe any of the statistics studied reflect the quality of an item adequately. We urge our colleagues to use the smoothed item characteristic curve to assess the measurement quality of items since it is the only satisfactory method available to us at this point in time.

Table 1.

Values Added to the In-fit for Differences in Student Achievement and the Item Calibration

RIT Difference	Logit Difference	Contribution If Answered Correctly	Contribution If Answered Incorrectly
-20	-2.0	.14	7.39
-19	-1.9	.15	6.69
-18	-1.8	.17	6.05
-17	-1.7	.18	5.47
-16	-1.6	.20	4.95
-15	-1.5	.22	4.48
-14	-1.4	.25	4.06
-13	-1.3	.27	3.67
-12	-1.2	.30	3.32
-11	-1.1	.33	3.00
-10	-1.0	.37	2.72
-9	-.9	.41	2.46
-8	-.8	.45	2.23
-7	-.7	.50	2.01
-6	-.6	.55	1.82
-5	-.5	.61	1.65
-4	-.4	.67	1.49
-3	-.3	.74	1.35
-2	-.2	.82	1.22
-1	-.1	.90	1.11
0	.0	1.00	1.00
1	.1	1.11	.90
2	.2	1.22	.82
3	.3	1.35	.74
4	.4	1.49	.67
5	.5	1.65	.61
6	.6	1.82	.55
7	.7	2.01	.50
8	.8	2.23	.45
9	.9	2.46	.41
10	1.0	2.72	.37
11	1.1	3.00	.33
12	1.2	3.32	.30
13	1.3	3.67	.27
14	1.4	4.06	.25
15	1.5	4.48	.22
16	1.6	4.95	.20
17	1.7	5.47	.18
18	1.8	6.05	.17
19	1.9	6.69	.15
20	2.0	7.39	.14

Table 2.

Comparison of Item Quality Statistics for Samples and the Total Population

	POINT BISERIAL				MEAN SQ IN-FIT				MEAN SQ OUT-FIT				F-COEFFICIENT			
TEM	TOT	LOW	HIGH	SAMP	TOT	LOW	HIGH	SAMP	TOT	LOW	HIGH	SAMP	TOT	LOW	HIGH	SAMP
2	0.41	0.23	0.13	0.46	1.0	1.1	1.1	0.9	1.0	1.2	1.6	0.6	0.99	0.77	0.75	0.98
4	0.53	0.31	0.36	0.57	0.9	0.9	0.9	0.8	2.5	1.0	1.0	0.9	0.99	0.95	0.93	0.96
6	0.39	0.37	0.10	0.43	0.8	0.9	1.1	0.9	1.5	1.1	1.4	0.7	0.97	0.84	0.63	0.95
12	0.32	0.30	-0.06	0.37	1.1	1.0	1.3	1.1	2.4	0.9	2.7	0.8	0.98	0.95	0.00	0.96
15	0.47	0.20	0.20	0.43	1.0	1.1	1.0	1.0	1.8	0.7	0.8	1.0	0.99	0.89	0.89	0.97
16	0.52	0.27	0.19	0.48	0.9	1.0	1.0	0.9	3.0	0.4	0.9	0.7	0.98	0.98	0.77	0.97
18	0.42	0.27	0.12	0.41	1.0	1.0	1.1	1.1	1.2	1.3	0.9	0.8	1.00	0.80	0.00	0.96
19	0.39	0.25	0.02	0.37	1.0	1.0	1.2	1.2	1.8	0.8	1.2	1.0	0.97	0.94	0.00	0.94
20	0.46	0.26	0.35	0.53	0.9	0.9	0.9	0.9	2.0	0.8	1.0	0.9	0.99	0.95	0.93	0.96
21	0.42	0.40	0.08	0.45	1.0	0.9	1.1	1.0	1.6	0.8	1.8	0.6	0.99	0.95	0.00	0.95
11*	0.17	0.12	0.19	0.23	1.4	1.1	1.0	1.3	5.9	1.2	0.5	1.0	0.63	0.82	0.90	0.83
25*	0.43	0.17	0.19	0.39	1.0	1.1	1.0	1.0	3.0	0.8	0.3	0.9	0.94	0.79	0.95	0.88
28*	0.33	0.03	0.21	0.30	1.2	1.2	1.0	1.2	5.5	2.5	1.3	1.5	0.92	0.00	0.73	0.76
38*	0.27	0.16	0.12	0.25	1.2	1.1	1.1	1.2	4.9	1.4	0.6	1.6	0.91	0.84	0.23	0.59
43*	0.32	0.14	0.14	0.30	1.2	1.1	1.0	1.3	4.0	0.8	1.1	1.4	0.89	0.85	0.68	0.86
54*	0.16	0.07	0.12	0.10	1.5	1.3	1.1	1.6	13.4	1.8	1.7	2.6	0.00	0.00	0.00	0.00

* Item with problem shown on item characteristic curve

Figure 1

Unsmoothed Item Curve

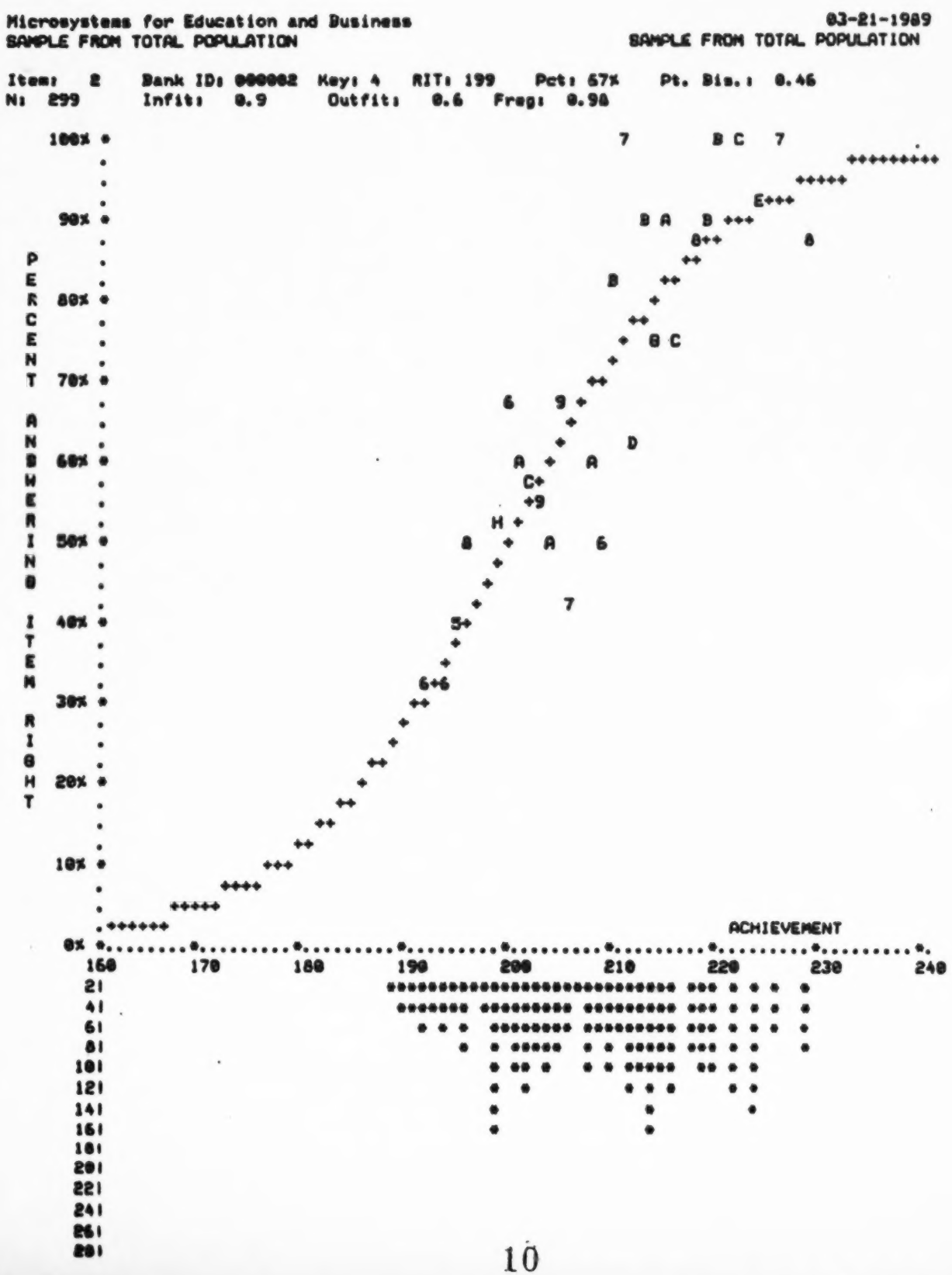

Figure 2

Smoothed Item Curve

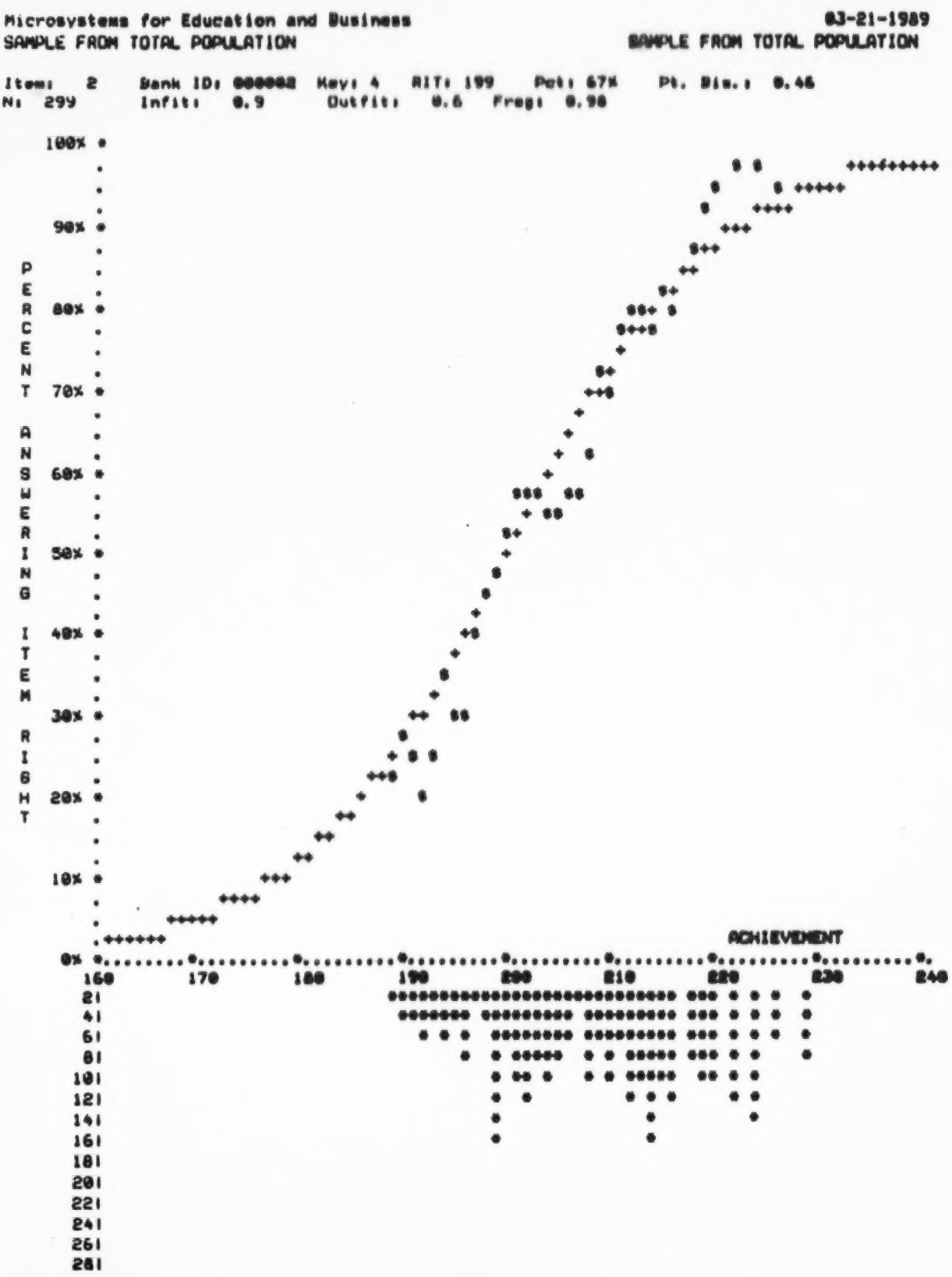

Figure 3

Example of an "Ideal" Item Curve

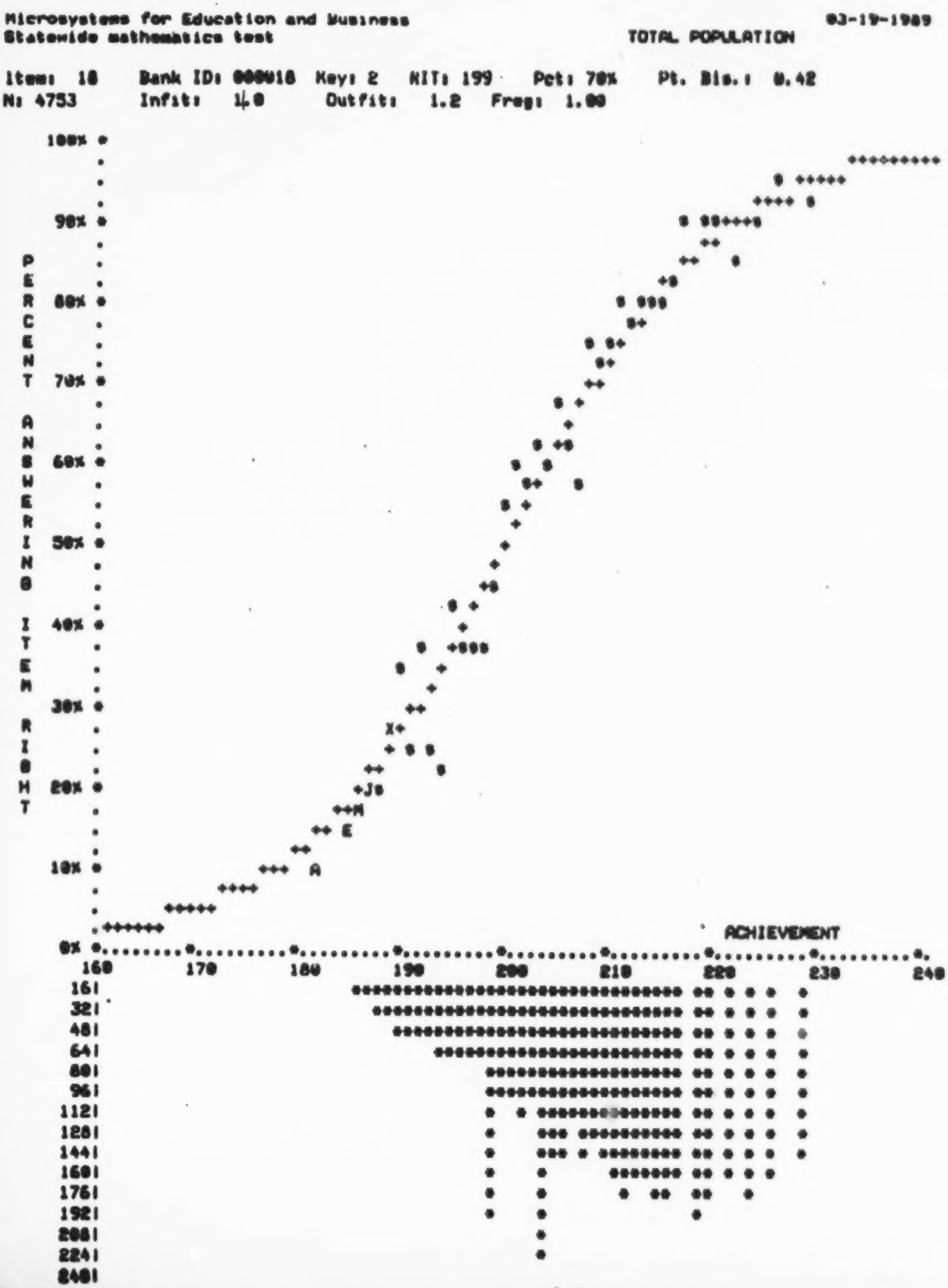

Figure 4

Example of a "Problem" Item Curve